Lecture 10: Endogeneity

Prepared for ECON 5033

Eric S. Lin

Department of Economics, National Tsing Hua University

December 3, 2024

Endogeneity Problem

- What is "endogeneity" problem?

$$\mathrm{E}[\varepsilon_i | X_i] \neq 0$$

- Recall that:

$$Y = X\beta + \varepsilon.$$

- The OLS estimator is NOT unbiased:

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &= \beta + \mathrm{E}[\left(X'X\right)^{-1} X'\varepsilon] = \beta + \mathrm{E}[\left(X'X\right)^{-1} X'\mathrm{E}[\varepsilon|X]] \\ &\neq \beta. \end{aligned}$$

- Also, OLS estimator is NOT consistent:

$$\begin{aligned} \mathrm{plim}\hat{\beta} &= \beta + \mathrm{plim}\left(\frac{X'X}{n}\right)^{-1} \mathrm{plim}\left(\frac{X'\varepsilon}{n}\right) \\ &= \beta + Q^{-1} \cdot \mathrm{plim}\left(\frac{X'\varepsilon}{n}\right) \neq \beta. \end{aligned}$$

Endogeneity Problem

- What is the source (type) of endogeneity?
  - omitted variables
  - measurement error
  - mutual causality or simultaneity
- What is the consequence of endogeneity?
  - estimation
  - testing

Omitted Variables

- Consider the simple linear regression model to estimate *return to schooling*,

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where

$$\begin{aligned} Y_i = &\ \text{log of wage} \\ X_i = &\ \text{years of education.} \end{aligned}$$

- It is reasonable to have,

$$\varepsilon_i = A_i + v_i.$$

- Unobserved heterogeneity (A_i)
- What is the covariance between X_i and ε_i?

Measurement Errors

- Consider the simple linear regression model.

$$Y_i = \alpha + \beta X_i^* + \varepsilon_i, \tag{1}$$

where

$$\begin{aligned} Y_i &= \text{log of wage} \\ X_i^* &= \text{intelligence.} \end{aligned}$$

- Since X_i^* is a latent variable we may use the IQ test to replace it empirically. However, (classical) "Measurement error" or "Error in variables" (EIV) issue may show up:

$$X_i = X_i^* + v_i,$$

where

$$\mathrm{E}[v_i | X_i^*] = 0, \ \mathrm{E}[v_i \varepsilon_i] = 0.$$

Measurement Errors

- Model (1) could be rewritten as

$$\begin{aligned} Y_i &= \alpha + \beta\left(X_i - v_i\right) + \varepsilon_i \\ &= \alpha + \beta X_i + u_i, \end{aligned}$$

where

$$u_i = \varepsilon_i - \beta v_i \quad (\text{composite errors})$$

- One could compute that,

$$\mathrm{E}[X_i u_i] = -\beta\sigma_v^2.$$

- In this case,

$$\text{plim}\hat{\beta} = \beta - \frac{\beta\sigma_v^2}{\sigma_X^2} = \beta\left(1 - \frac{\sigma_v^2}{\sigma_{X^*}^2 + \sigma_v^2}\right) = \frac{\beta}{1+\lambda},$$

where

$$\lambda = \frac{\sigma_v^2}{\sigma_{X^*}^2} \geq 0.$$

Measurement Errors

- Note that:

$$\text{plim}\hat{\beta} = \frac{\beta}{1+\lambda}, \ \lambda = \frac{\sigma_v^2}{\sigma_{X^*}^2} \geq 0$$

- What's the direction of the bias?
  - "Bias toward zero", "attenuation"
- What if for multi-covariates case? Coefficient without measurement error? What if for several EIVs?
- EIV is common in micro data, e.g., household income survey, lottery winners
- The magnitude of the bias does not depend on the magnitude of σ_v^2 but on the "noise to signal" ratio.
  - Interpret $\lambda = 1$ as a special case?

Simultaneous Equation Models

- Consider the Simultaneous Equation Models (SEM) in demand and supply systems.

$$
\begin{aligned}
Q_i^D &= P_i\beta_D + X_{Di}'\gamma_D + \varepsilon_{Di} \\
Q_i^S &= P_i\beta_S + X_{Si}'\gamma_S + \varepsilon_{Si} \\
Q_i^D &= Q_i^S,
\end{aligned}
$$

- Structural form equations
  - Behavioural equations
- Endogenous variables: P_i and Q_i
- Equilibrium condition
- Exogenous variables: X_{Di}' and X_{Si}'
- Structural form parameters

Simultaneity Problem

- One could express the endogenous variables in terms of exogenous variables:

$$P_i = \frac{1}{\beta_D - \beta_S}\left(-X'_{Di}\gamma_D + X'_{Si}\gamma_S - \varepsilon_{Di} + \varepsilon_{Si}\right) \tag{2}$$

$$Q_i = \frac{\beta_D}{\beta_D - \beta_S}\left(-X'_{Di}\gamma_D + X'_{Si}\gamma_S - \varepsilon_{Di} + \varepsilon_{Si}\right) + X'_{Di}\gamma_D + \varepsilon_{Di}. \tag{3}$$

- Reduced form equations (parameters)
- Can we do OLS estimators of regressing P_i on (X'_{Di}, X'_{Si}) and Q_i on (X'_{Di}, X'_{Si})?
  - obtaining RF coefficients?
- How about estimating SF coefficients?
  - we say there is endogeneity or simultaneity problem.

A Simple SEM

- Most popular two-equation SEM is given by,

$$\begin{aligned} y_i &= \gamma' Y_i + \beta' X_{1i} + \varepsilon_i \\ &= \begin{pmatrix} Y_i \\ X_{1i} \end{pmatrix}' \begin{pmatrix} \gamma \\ \beta \end{pmatrix} + \varepsilon_i \\ &= Z_i' \delta + \varepsilon_i, \end{aligned}$$

who the fuck is Z exactly from now on? No introduction?
So Z includes both?

where it is known or suspected that,

$$\begin{aligned} \mathrm{E}\left[Y_i \varepsilon_i\right] &\neq 0 \\ \mathrm{E}\left[X_{1i} \varepsilon_i\right] &= 0. \end{aligned}$$

- Endogenous variables
- Exogenous variables
  - included exogenous variables

A Simple SEM

- Now assume that we have a reduced form for Y_i given by,

$$\begin{aligned} Y_{i(G\times 1)} &= \Pi_1' X_{1i(k_1\times 1)} + \Pi_2' X_{2i(k_2\times 1)} + V_i \\ &= \Pi' X_i + V_i. \end{aligned}$$

- For the jth element of Y_i we have,

$$Y_{ji} = \Pi_{j1}' X_{1i} + \Pi_{j2}' X_{2i} + V_{ji}.$$

- We will assume that,

$$\begin{aligned} \mathrm{E}\left[X_{1i} V_{ji}\right] &= 0 \\ \mathrm{E}\left[X_{2i} V_{ji}\right] &= 0. \end{aligned}$$

- What is X_{2i}? excluded exogenous variables
- Properties of OLS in estimating the RF? Why is it unfortunate?

A Simple SEM

- Consider again the two-equation model in reduced form.

$$\begin{aligned} y_i &= (\Pi_1\gamma + \beta)' X_{1i} + (\Pi_2\gamma)' X_{2i} + \varepsilon_i + \gamma' V_i \\ &= \pi_1' X_{1i} + \pi_2' X_{2i} + v_i \\ Y_i &= \Pi_1' X_{1i} + \Pi_2' X_{2i} + V_i, \text{ where} \\ \pi_1 &= \Pi_1\gamma + \beta \\ \pi_2 &= \Pi_2\gamma. \end{aligned}$$

- If all variables are scalars, we have that,

$$\begin{aligned} y_i &= \gamma Y_i + \beta X_{1i} + \varepsilon_i \\ Y_i &= \Pi_1 X_{1i} + \Pi_2 X_{2i} + V_i, \text{ where} \\ \pi_1 &= \Pi_1\gamma + \beta \\ \pi_2 &= \Pi_2\gamma. \end{aligned} \tag{4}$$

A Simple SEM

- Consider the exactly identified case.
- Can write structural parameters as functions of RF parameters
- This suggests an estimator that involves the sample analogs of the reduced form parameters. Let,

$$\begin{pmatrix} \hat{\Pi}_1 \\ \hat{\Pi}_2 \end{pmatrix} = \left(X'X\right)^{-1} X'Y$$

$$\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} = \left(X'X\right)^{-1} X'y.$$

- The Indirect Least Squares Estimator (ILS) is given by,

$$\hat{\gamma} = \hat{\Pi}_2^{-1}\hat{\pi}_2, \; \hat{\beta} = \hat{\pi}_1 - \hat{\Pi}_1\hat{\Pi}_2^{-1}\hat{\pi}_2.$$

- One can actually show that $(\hat{\gamma} : \hat{\beta})' = \left(X'Z\right)^{-1} X'y$.

A Simple SEM

Theorem

Given $K_2 = G$ (exactly identified), let $\hat{\hat{\delta}} = (\hat{\gamma} : \hat{\beta})'$. The ILS estimator of δ is $\hat{\hat{\delta}} = (X'Z)^{-1} X'y$.

Proof.

The relationship between structural and reduced form parameters satisfies,

$$\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} = \begin{pmatrix} \hat{\Pi}_1 & I \\ \hat{\Pi}_2 & 0 \end{pmatrix} \begin{pmatrix} \hat{\gamma} \\ \hat{\beta} \end{pmatrix}.$$

Use the fact that

$$(X'X)^{-1} X'y = (X'X)^{-1} X' (Y : X_1) \hat{\hat{\delta}} = (X'X)^{-1} X'Z\hat{\hat{\delta}}.$$

Canceling out the term $(X'X)^{-1}$ both sides leads to the result. □

Why We Need IVs?

- Consider the following model:

$$y_i = \gamma Y_i + \varepsilon_i$$

- Is γ the marginal effect of Y_i on y_i?

$$\frac{\partial y_i}{\partial Y_i} \stackrel{?}{=} \gamma$$

- If $\text{cov}(Y_i, \varepsilon_i) \neq 0$

$$\frac{\partial y_i}{\partial Y_i} = \gamma + \frac{\partial \varepsilon_i}{\partial Y_i} \neq \gamma$$

- marginal effect can be interpreted as usual
- Idea is to isolate the part of Y_i which is related to ε_i
- Use the uncorrelated part of Y_i and ε_i to estimate the marginal effect

IV Conditions

- The estimator $\hat{\delta}$ is also called Instrumental Variable estimator (IV).
- Generally, a good instrument should satisfy three conditions
  1. $\mathrm{E}[X_i \varepsilon_i] = 0$. [exogeneity condition]
  2. $\mathrm{E}[X_i' Z_i] \neq 0$. [relevance condition]
  3. $\mathrm{E}[X_i' Z_i]$ should not be small. [not weak relevance]
- The complete set of instruments is $X = (X_1 : X_2)$
- Instrument for X_1? Instrument for Y?
- The sample counterpart of condition 1: $\mathrm{E}[X_i \varepsilon_i] = 0$ is,

$$\frac{1}{n} \sum_{i=1}^{n} X_i \left(y_i - Z_i' \delta \right) = \frac{1}{n} X' \left(y - Z\delta \right). \tag{5}$$

- The IV estimator is to find the value of δ that makes (5) equal zero.
- Note that we are using X_2 as an instrument for Y and so that the order condition is that there be as many instruments (not including anything in X_1) as there are offending variables.

IV

- Consider the very simple example.

$$\begin{aligned} y_i &= \gamma Y_i + \varepsilon_i \\ Y_i &= \Pi_2 X_{2i} + V_i, \end{aligned}$$

where all are scalars.

- The IV estimator has the form,

$$\hat{\gamma} = \frac{\sum_{i=1}^n X_{2i} y_i}{\sum_{i=1}^n X_{2i} Y_i} = \gamma + \frac{\sum_{i=1}^n X_{2i} \varepsilon_i}{\sum_{i=1}^n X_{2i} Y_i}.$$

- Using the fact,

$$\begin{aligned} \frac{\sum_{i=1}^n X_{2i} \varepsilon_i}{n} &\xrightarrow{p} \mathrm{E}\left[X_{2i} \varepsilon_i\right] = 0 \\ \frac{\sum_{i=1}^n X_{2i} Y_i}{n} &\xrightarrow{p} \mathrm{E}\left[X_{2i} Y_i\right] = \Pi_2 \mathrm{E}[X_{2i}^2] \neq 0, \end{aligned}$$

IV

- One would show,

$$\hat{\gamma} \xrightarrow{p} \gamma.$$

- There are two cases where the consistency of IV estimator fails.
  1. Irrelevant instrument – $\Pi_2 = 0$. Rank condition fails.
  2. Invalid instrument – $\mathrm{E}[X_{2i}\varepsilon_i] \neq 0$. We omit a relevant variable.
- We've demonstrated that indirect least squares will be useful for exactly identified simultaneous model.
- Recall that the IV or ILS estimator has the form,

$$\hat{\delta}_{IV} = \left(X'Z\right)^{-1} X'y,$$

where

$$Z = \left(Y : X_1\right).$$

IV

- Consistency: $\hat{\delta}_{IV} \xrightarrow{p} \delta$
- The asymptotic distribution of IV estimator $\hat{\delta}_{IV}$ is,

$$\sqrt{n}(\hat{\delta}_{IV} - \delta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \text{plim}\left(\frac{X'Z}{n}\right)^{-1}\left(\frac{X'X}{n}\right)\left(\frac{Z'X}{n}\right)^{-1}\right).$$

IV

- Note that,

$$
\begin{aligned}
&\operatorname{plim}\left(\frac{X'Z}{n}\right)^{-1}\left(\frac{X'X}{n}\right)\left(\frac{Z'X}{n}\right)^{-1} \\
&= \operatorname{plim}\left[Z'X\left(X'X\right)^{-1}\frac{X'X}{n}\left(X'X\right)^{-1}X'Z\right]^{-1} \\
&= \operatorname{plim}\left[\begin{pmatrix} \hat{\Pi}_1 & I \\ \hat{\Pi}_2 & 0 \end{pmatrix}'\frac{X'X}{n}\begin{pmatrix} \hat{\Pi}_1 & I \\ \hat{\Pi}_2 & 0 \end{pmatrix}\right]^{-1} \\
&= \left[\begin{pmatrix} \Pi_1 & I \\ \Pi_2 & 0 \end{pmatrix}'Q\begin{pmatrix} \Pi_1 & I \\ \Pi_2 & 0 \end{pmatrix}\right]^{-1}.
\end{aligned}
$$

IV

- Inference? Estimated variance matrix,

$$
\begin{aligned}
& s^2 \left(X'Z\right)^{-1} \left(X'X\right) \left(Z'X\right)^{-1} \\
& = s^2[Z'X \left(X'X\right)^{-1} X'Z]^{-1} \\
& = s^2 \left(Z'P_X Z\right)^{-1},
\end{aligned}
$$

where

$$
s^2 = \frac{1}{n}(y - Z\hat{\delta}_{IV})'(y - Z\hat{\delta}_{IV}).
$$

- The usual t, F tests can apply.

IV vs. 2SLS

- The IV or ILS estimators could be implemented practically using a two-step procedure.

1. Regress Y_i on X_i and obtain the predictions $\hat{Y}_i$.
   this is the part of Y_i which is uncorrelated with ε_i
2. Regress y_i on $\hat{Y}_i$ and X_{1i} to get $\hat{\delta}_{2SLS}$.
   the new regressor $\hat{Y}_i$ is no longer correlated with ε_i

- The estimator is known as two stage least squares estimator (2SLS or TSLS).
- One could formally prove that 2SLS is equivalent to IV estimator.

IV, 2SLS & ILS

Theorem

Under exact identification, 2SLS = IV = ILS.

Proof.

We have,

$$\hat{Y} = P_X Y$$
$$\hat{X}_1 = P_X X_1 = X_1$$
$$\hat{Z} = (\hat{Y} \vdots X_1) = (\hat{Y} \vdots \hat{X}_1) = P_X Z.$$

2SLS is,

X'Z must be invertible

$$\hat{\delta}_{2SLS} = \left(Z' P_X Z\right)^{-1} Z' P_X y = \left(X' Z\right)^{-1} X' y = \hat{\delta}_{IV}.$$

Overidentified Model

- If the model is overidentified, what's wrong with the IV estimator?
- More instruments than we need because $k_2 > G$.
- $\hat{\pi}_2 = \hat{\Pi}_2 \hat{\gamma}$ does not have a unique solution.
  - $k_2 \times G$
  - $\hat{\delta}_{2SLS} = (Z' P_X Z)^{-1} Z' P_X y \stackrel{?}{=} (X'Z)^{-1} X'y = \hat{\delta}_{IV}$
- Several methods:
  - 2SLS
  - IV
  - GMM

Overidentified Model – 2SLS

- Recall that the two stage procedure.
  1. Regress Z on X and obtain the prediction $\hat{Z}$.
  2. Regress y on $\hat{Z}$ and the estimates will be the 2SLS estimates.
- We get an estimator of the form,

$$\hat{\delta}_{2SLS} = (Z'X\left(X'X\right)^{-1}X'Z)^{-1}Z'X\left(X'X\right)^{-1}X'y$$

- However, we do not have the following form:

$$\hat{\delta}_{2SLS} = \left(X'Z\right)^{-1}X'y,$$

where the IV is X

  - can we still use this form?

Overidentified Model – IV

- In other words, we treat $\hat{Z}$ as IV.

$$\hat{\delta}_{IV} = (\hat{Z}'Z)^{-1}\hat{Z}'y = (\hat{Z}'\hat{Z})^{-1}\hat{Z}'y = \hat{\delta}_{2SLS}.$$

- The instrument is,

$$X\left(X'X\right)^{-1}X'Z = (X\left(X'X\right)^{-1}X'Y,\ X_1).$$

- Interpretation? $X\left(X'X\right)^{-1}X'Y$, X_1
- Asymptotic distribution of 2SLS:

$$\sqrt{n}(\hat{\delta}_{2SLS} - \delta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\left[\begin{pmatrix} \Pi_1 & I \\ \Pi_2 & 0 \end{pmatrix}' Q \begin{pmatrix} \Pi_1 & I \\ \Pi_2 & 0 \end{pmatrix}\right]^{-1}\right).$$

2SLS Standard Errors

- Issue on the estimated variance covariance matrix of the 2SLS estimator.

$$\widehat{\text{var}}[\hat{\delta}_{2SLS}] = \hat{\sigma}^2 (Z'X \left(X'X\right)^{-1} X'Z)^{-1}$$
$$\hat{\sigma}^2 = \frac{1}{n}(y - Z\hat{\delta}_{2SLS})'(y - Z\hat{\delta}_{2SLS}).$$

- Something wrong from the 2-stage procedure described above, why?
  - regression implemented, false variance matrix estimate and true regression

$$y = \hat{Z}\delta + error$$
$$\frac{1}{n}(y - \hat{Z}\hat{\delta}_{2SLS})'(y - \hat{Z}\hat{\delta}_{2SLS})$$
$$y = \hat{Z}\delta + (Z\delta - \hat{Z}\delta) + \varepsilon$$

2SLS Standard Errors

- Another issue on the 2SLS standard error
- The 2SLS standard error is typically quite high compared to that of the OLS estimator
- The most important reason for this is that instrument and regressor have a low correlation
- Note that condition 3: $E[X_i' Z_i]$ should not be small
- Need to take care of "weak instrument" or "weak IV" problem!

GMM

- Conditional moment restriction,

$$\mathrm{E}\left[\varepsilon_i | X_i\right] = 0$$

- Unconditional moment restriction,

$$\mathrm{E}\left[X_i \varepsilon_i\right] = 0 \tag{6}$$

- The sample counterpart of the left hand side in (6) is,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \left(y_i - Z_i'\delta\right) = \frac{1}{n} X' \left(y - Z\delta\right).$$

We cannot make this exactly zero in general except the just identified case.

GMM

- Instead we make the quadratic form close to zero by solving,

$$\min Q(y,\delta) = \min_{\delta} (y - Z\delta)' XX' (y - Z\delta)$$

- However, the following is more general:

$$\min Q(y,\delta) = \min_{\delta} (y - Z\delta)' X W_n X' (y - Z\delta)$$

for some positive definite weighting matrix W_n.

  - non-negative measure
  - objective function is zero if the $X'(y - Z\delta)$ is zero

- Solution?

$$\hat{\delta}_{GMM}(W_n) = \left(Z'XW_nX'Z\right)^{-1} Z'XW_nX'y.$$

GMM

- Recall the general GMM estimator

$$\hat{\delta}_{GMM}(W_n) = \left(Z'XW_nX'Z\right)^{-1} Z'XW_nX'y.$$

- This is a consistent estimator regardless the choice of W_n!
- Nevertheless, we can make it more efficient by choosing W_n properly
- Best (optimal) W_n?

$$W_n = [\text{var}[X'\varepsilon]]^{-1} = \sigma^{-2}\left(X'X\right)^{-1}$$

    - Intuition?
- What does $\hat{\delta}_{GMM}(W^*)$ look like?

GMM – Summary

- For overidentified case, try to minimize $Q(y, \delta)$
- For exactly identified case, $Q(y, \delta) = 0$ all the time
- For exactly identified case, choice of weighting matrix does not matter
  - since $Q(y, \delta) = 0$ all the time
- Sometimes 2SLS, IV and GMM estimators are called generalized instrumental variables estimator (GIVE).
- If the error terms are heteroskedastic or autocorrelated, the optimal weighting matrix should be adjusted accordingly.

Two Types of Tests

- Test if you have endogeneity regressor or not
  - endogeneity test
  - exogeneity test
  - Hausman test or Hausman's specification test
  - Durbin-Wu-Hausman test
- (partially) Test if your IV is really exogenous or not
  - IV exogeneity test
  - IV endogeneity test
  - Sargan test
  - Hansen's $\mathcal{J}$ test
  - Over-identification test
  - Over-identifying restriction test
  - Model specification test

Hausman Test

- A.k.a the Hausman's specification test
- Consider two estimators: efficient estimator (OLS) and consistent estimator (IV)
- If no endogeneity, two estimators are sufficiently close
- Given endogeneity, the efficient estimator is most likely inconsistent
- Under the null,
  - $\hat{\beta}_{OLS}$ will be consistent and efficient
  - $\hat{\beta}_{IV}$ will be consistent as well but not efficient.
- Under the alternative hypothesis
  - $\hat{\beta}_{OLS}$ is inconsistent
  - $\hat{\beta}_{IV}$ is still consistent
- The Hausman test statistic is given by,

$$\mathcal{H} = (\hat{\beta}_{OLS} - \hat{\beta}_{IV})'[\text{var}[\hat{\beta}_{IV} - \hat{\beta}_{OLS}]]^{-1}(\hat{\beta}_{OLS} - \hat{\beta}_{IV}) \sim \mathcal{X}^2(G)$$

  - Problem?

Hausman Test

Lemma

When the null is true, the following relationship holds.

$$cov(\hat{\beta}_{OLS}, \hat{\beta}_{IV}) = var[\hat{\beta}_{OLS}].$$

- Can simplify the covariance matrix in the Hausman test statistic.
- Test the equivalence of OLS and IV estimators through testing:
  - one of the coefficient by a t test
  - all the parameters by a $\mathcal{X}^2$ test
- G equals the number of potentially endogenous regressors
- The test statistic is now given by,

$$\mathcal{H} = (\hat{\beta}_{OLS} - \hat{\beta}_{IV})'[\text{var}[\hat{\beta}_{IV}] - \text{var}[\hat{\beta}_{OLS}]]^{-1}(\hat{\beta}_{OLS} - \hat{\beta}_{IV}) \sim \mathcal{X}^2(G)$$

Durbin-Wu-Hausman Test

- A computationally equivalent way of testing the same hypothesis is to estimate the following auxiliary regression by OLS:

$$y = \gamma Y + \beta X_1 + \theta e_2 + \nu,$$

  where e_2 is the residual from the reduced form equation for Y.
- Test $H_o : \theta = 0$ – Durbin-Wu-Hausman test
- This reproduces the IV estimator for γ and β
- Heteroskedasticity robust s.e.
- "Control function" approach

Test for IV Exogeneity

what the f is "IV estimation", new term?

- Whether the instruments are exogenous is important.

$$E[X_i \varepsilon_i] = 0.$$

- It is needed to have a test for the validity of instruments.
  - If IVs are not valid, Hausman test is not correct anymore.
- Intuitive way is to regress ε_i on X_i but ...
- Sargan test on the validity of instruments can be computed as follows.
  1. Apply the IV estimation to the level regression and obtain residual, $\hat{\varepsilon}_{IV}$
  2. Perform the regression $\hat{\varepsilon}_{IV} = X\theta + v$
  3. Compute LM $= nR^2$ of the regression in step 2. Under the null that the instruments are exogenous, the LM has an asymptotic $\mathcal{X}^2(k_2 - G)$ distribution.

Test for Overidentification

- Consider the situations that there are more instruments than endogenous variables.
- This could be tested by employing the objective function used to find 2SLS or IV, i.e., GMM criterion function
- The null is

$$\mathrm{E}[X_i \varepsilon_i] = 0$$

- Hasen's (1982) $\mathcal{J}$ statistic:

$$\mathcal{J} = \frac{1}{\hat{\sigma}^2}(y - Z\hat{\delta}_{IV})' X(X'X)^{-1}X'(y - Z\hat{\delta}_{IV}) = nR^2.$$

I thought you said Z includes both type, so where does delta IV come from?

Test for Overidentification

- Limiting distribution of the test statistic:

$$\mathcal{J} \xrightarrow{d} \mathcal{X}^2(k_2 - G)$$

  - One may test the hypothesis by comparing the $\mathcal{J}$ statistic to a $\mathcal{X}^2$ critical value
  - Small $\mathcal{J}$ value will accept the null
    - unusual compared to a typical hypothesis testing
  - What if for exactly identified model?

My Own Research Related to Endogeneity

- Instructors teaching effectiveness vs. research capability
- Math test scores in junior high vs. cramming activity
- Military spending vs. unemployment/ inequality/ welfare spending
- Food prices vs. civil war/ international war
- Production/process innovation vs. firm performance
- County-level divorce rates vs. fertility rate and FLPR
- County-level poverty rates vs. unemployment rate
- MLB team performance vs. wage inequality, attendance
- Admission method vs. NTHU students performance
- THC residents vs. academic/non-academic performance
- PhD ETTD vs. financial aid
- Postdoctoral experience vs. academic career job choice